Fermi National Accelerator Laboratory

FERMILAB-Conf-92/277

Experiences with the ACPMAPS 50 GFLOP System

Mark Fischler

Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510

October 1992

Presented at *Computers in High Energy Physics*,
Annecy, France, September 21-25, 1992

Disclaimer

Experiences With the ACPMAPS 50 GFLOP System

Mark Fischler
Fermilab, Batavia IL 60510 USA

The Fermilab Computer R&D and Theory departments have for several years collaborated on a multi-GFLOP (recently upgraded to 50 GFLOP) system for lattice gauge calculations. The primary emphasis is on flexibility and ease of algorithm development. This system (ACPMAPS) has been in use for some time, allowing theorists to produce QCD results with relevance for the analysis of experimental data. We present general observations about benefits of such a scientist-oriented system, and summarize some of the advances recently made. We also discuss what was discovered about features needed in a useful alogrithm exploration platform. These lessons can be applied to the design and evaluation of future massively parallel systems (commercial or otherwise).

1. Introduction — The ACPMAPS Project

The ACPMAPS project at Fermilab, initiated in 1987 as a collaborative effort between the lattice gauge theorists and the Advanced Computer Program, was created to provide the theorists with computational power to do useful QCD calculations. At the time, several special-purpose efforts in this field were in various stages of startup or production, notably the series of machines at Columbia, the GF-11, the APE project in Europe, and QCDPAX in Japan. It was felt at Fermilab that the important need we could fulfill was for a powerful flexible system, on which complex algorithms could quickly and easily be brought up. Since algorithm advances were (and still are) as important as advances in CPU power, a massively parallel system suitable for studying many approaches at some reasonable efficiency would be a valuable asset.

Creating a machine for algorithm exploration sets goals for the nature of the system and software support. For example, since not all methods can be cast in a SIMD or lock-step communications mold, the machine must be MIMD. The scientific users should not be forced to become experts in massively parallel computing, or in the particular machine architecture — this implies that the programming paradigm must shield the user from the details of the system architecture. The coding tools must allow the user to express algorithms in terms familiar to the field of interest. To maximize the usefulness of the system, smooth multi-user sharing and appropriate massively parallel I/O must be supported.

To approach the software goals, we have created a concept oriented tool set for coding grid-like problems, called CANOPY [1]. CANOPY deals with concepts such as grids (with particular connectivities), sites on the grids, field data associated with the sites, and tasks to be done for some set of sites on a grid. The user describes the algorithm in terms of the task done at a single site, employing such concepts as obtaining field data from a particular neighboring site. Concepts supported include link fields, multiple grids, and maps from one grid to another. The user is shielded from such issues as how data and work are distributed, and how to get field data which may reside on remote nodes. The CANOPY program can be run on arbitrary numbers of processor nodes, and can trivially be moved to any system supporting the tool set.

The nature of a given system architecture will determine if one can create such tools, and ensure that their use will not entail severe efficiency penalties. The natural strategy is that each site, and its associated field data, be treated as a "virtual processor"; since the natural granularity of the code is the work done at a single site, CANOPY programs will tend to do frequent small data transfers. The architectural requirements for this strategy are MIMD processing, "flat" global communication (any node can access data on any other node), and reasonably low communications overhead and latency.

The ACPMAPS system (described more fully in [2]) is based on a backbone of crossbar switching crates. The processors reside in slots in these active-backbone crates, and can access nodes in their own crate, or in remote crates via intercrate cables (the switching crates have routing information). Thus the communications system is analagous to a telephone network, with reconfiguration time on the order of a microsecond, and channel bandwidths of 20 Mbyes/second. The 5 GFLOP first generation system (in use since 1989) has processor nodes based on the 20 MFLOP Weitek 8032 chip set. We are currently moving to more powerful nodes, based on pairs of 80 MFLOP Intel i860's. For the past year, new modules totalling 20

GFLOPs in peak power have been inserted into unused slots in the system and a test stand (and are being used for some physics calculations). Until recently, the physics being done dictated that we not risk removing the Weitek nodes; the cut-over has now been made to a 50 GFLOP system based wholely on i860's (existing CANOPY applications run on the new processors without conversion).

2. Physics Results and Directions

The mainstream lattice gauge theory efforts on ACPMAPS over the past few years have focussed on the physics of systems containing one or more heavy quarks, on charmiomium spectroscopy, and on exploration of improved algorithms for handling fermions. Many of these efforts required exploration of multiple, complex algorithms, or inclusion of terms which might be burdensome to code on less flexible systems, to explore systematic errors and implement methods to reduce these errors.

For example, several methods of including in propagator calculations terms in the fermion action at higher-order in the lattice spacing were investigated. Including these terms can bring the finite lattice spacing error in the fermionic action to order a^2 (matching the error in the pure gauge action). This was found to be a worthwhile improvement [3]. While these higher order terms can be implemented on other systems, CANOPY made it easy and routine to try a variety of changes, thus facilitating the exploration. Another improvement facilitiated by the flexible system was in using multistate operator smearing techniques to more accurately extract physics measurables from lattice configurations [4].

Perhaps the most published important results to date involve the lattice extraction of α_S and $\Lambda_{\overline{MS}}$ from charmonium spectoscopy. These computations were done with reliable estimates of all systematic and statistical uncertainties, including the effect of ignoring closed quark loops [5]. Thus, these results can be compared directly with experimental data, to look for deviations from QCD (they agree fairly well). At the recent LAT92 conference on Lattice Gauge Theory, results were presented on the spectrum of mesons in the heavy-light limit, f_B dependance on light quark mass and on lattice volume and spacing, spin splittings in charmonium, and spectroscopy of excited states, including the first direct measurement of radially excited states.

It is obvious that understanding and reliably estimating systematic uncertainties in lattice calculations is important. What we are observing is that addressing the issue of knowing these uncertainties often makes a big difference in what sort of programs need to be run, and in the nature of an appropriate environment for doing the research. This difference argues for more flexibility and ease of coding, and larger memory space, even at the expense of less raw compute power.

From the machine architecture standpoint, the key feature shared by most of the lattice gauge applications is that they are "tightly coupled" — the internode communication bandwidth and latency are important. For the 50 GFLOP ACPMAPS upgrade, the ten-fold increase in computational power is not accompanied by an upgrade of the communication subsystem. Thus for many problems, the new system will be limited by communication rather than CPU power. We are currently implementing techniques to partially alleviate this effect, by automatically coalescing transfers, so as to minimize the overhead encountered while communications resources are in use.

Side applications which have been run on ACPMAPS (typically making use of the new i860 nodes before the lattice gauge applications were ready to) include a Monte-Carlo integration for Jet Physics [6] applicable for analysis of CDF data. The uses outside of Lattice Gauge Theory tend to be more loosely coupled.

3. Concepts Applicable to Massively Parallel Systems for HEP

The goals necessary to create a good algorithm exploration platform — MIMD, coding tools to facilitate algorithm expression and shield the user from the details of machine parallellism, good system sharing and parallel I/O — are certainly useful features beyond the field of Lattice Gauge Theory. Other high energy physics users are demonstrating that a system like ACPMAPS can be useful (though not essential) to them. The lesson to be learned is **not** that physicists should continue to build our own systems to get the features we want. However, when systems designed to support tightly coupled applications become commercially available from several

vendors, the communications fabric may add only 10-15% to the cost of the overall system. In that case, there is a considerable advantage to acquiring a system, which can support uses ranging from theory calculations, to event reconstruction, to physics analysis of reconstructed events, to real-time data filtering. Even when the communications capabilities are not manifestly needed, they may lighten some problems (for example, the issue of how to break up farms of processors to avoid bottlenecks may become trivial), and the advantages in terms of supporting one system rather than several kinds of systems can be large. Assuming companies solve the communications issues at a reasonable cost, this may portend a movement back away from special-purpose systems.

When a tightly coupled architecture is applied to a loosely coupled application (such as event reconstruction), there is no problem with communication. However, there is no guaranteee that the same paradigm and tool sets will be optimal. Although existing non-lattice work on ACPMAPS has utilized the CANOPY software, the CPS Cooperative Processes Software developed for the Fermilab event reconstruction farms might well support more natural concepts for these problems. One might approach this issue by attempting to unify the tool sets into one larger set, but we feel that the proper approach is that of concept domain tool sets. That is, packages applicable to entire domains of problems — grid-like, manifest loose coupling, particle/cell, and so forth — are created. These are based on a standardized underlying model of the communications architecture of a system, so that the tool sets can be implemented on a wide variety of suitable platforms. This strategy encourages the user to absorb only those concepts which will be useful for relevant domains, and allows the individual tool sets to be manageable in scope.

Commercial systems with excellent communications fabrics, good compiler/system software, and massive I/O capabilities will be emerging in the near future — several of the various proposed "TeraFLOP" systems fit this description. The physics community can take fullest advantage of such machines working with industry to ensure the suitability of these systems, and by creating and supporting concept oriented tool sets that allow most researchers to focus on the science, rather than on the computer aspects of their applications.

References

1 Details of CANOPY can be found in the Canopy 5.0 Manual, M. Fischler, G. Hockney, P. Mackenzie, available from the Fermilab Computing Division.

2 The ACPMAPS System — A Detailed Overview, M. Fischler, FERMILAB-TM-1780.

3 A. El-Khadra, Nucl. Phys B (Proc. Supp.) 26 (1992), 372.

4 A. Duncan, E. Eichten, G. Hockney, and H. Thacker, Nucl. Phys B (Proc. Supp.) 26, 391 (1992).

5 A. El-Khadra, G. Hockney, A. Kronfeld, P. Mackenzie, Phys. Rev. Lett. 69, 729 (1992).

6 W. Giele, E. Glover, D. Kosower, Fermilab preprint FNAL-PUB-92-230-T.